Agentic NLP for Thematic Analysis of Fed Speeches: Yield Impacts under Jerome Powell (2018–2025)

Abstract. This study investigates whether the thematic emphasis in speeches during Federal Reserve Chair Jerome Powell's tenure influenced U.S. Treasury yields. We construct a novel dataset by scraping over 1,200 speeches from the Fed's official archive and tagging individual paragraphs using a LangGraph pipeline orchestrating LLaMA 3.1-based agents. Paragraphs are classified into inflation, employment, or other themes, and speech-level emphasis scores are calculated. To address multicollinearity in compositional data, we use only inflation and employment emphasis in the final model. The dataset is restricted to Powell's tenure (2018–2025) to control for leadership-driven variability, allowing a focused analysis on a consistent communication style and policy regime. An event-study framework is applied to detect abnormal yield movements around speech dates. While the federal funds rate significantly predicts yield changes, emphasis scores show no direct statistical impact. We discuss implications for NLP in economic analysis and propose future improvements. All code and data are open-sourced, and the results contribute to broader discussions at the intersection of AI and macro-finance.

1 Introduction

Federal Reserve communication has undergone a significant transformation over the past four decades. In the 20th century, monetary policy decisions were often opaque, with minimal forward guidance. This began to shift under Chair Alan Greenspan, who introduced measured signals to guide market expectations. Under Ben Bernanke and Janet Yellen, the Fed formalized forward guidance and increased transparency, releasing meeting minutes and projections to anchor expectations (1). Chair Jerome Powell's tenure (2018–2025) marks a continuation of transparency, but with a stronger reliance on narrative framing, often tailored to evolving labor market and inflation conditions. This evolution sets the stage for applying modern AI techniques to uncover the informational value embedded in central bank speeches.

Literature Review. A growing body of research explores the relationship between central bank communication and financial markets. Early work relied on dictionary-based sentiment scoring of earnings calls and regulatory filings, such as the Loughran-McDonald dictionary tailored to financial texts (6). Subsequent studies applied topic modeling and textual regressions to FOMC minutes and policy speeches to capture the thematic evolution of central bank narratives (5). More recent developments have used transformer-based models to detect changes in tone and transparency in ECB and Federal Reserve communication (4), and to link central bank sentiment to short-term equity volatility and credit spreads.

However, these approaches often suffer from key limitations. Many rely on bag-of-words or TF-IDF representations, which ignore word order, contextual meaning, and structural cues. Document-level aggregation blurs intra-textual variation, making it difficult to isolate fine-grained signals. In particular, most existing NLP pipelines lack paragraph-level granularity and do not harness the compositional or state-aware reasoning capabilities of modern large language models (LLMs). Furthermore, few studies operationalize communication as a time-localized signal within an event-study design, especially in the context of long-term fixed income markets such as Treasury yields.

This paper builds on these strands of literature by introducing an agentic NLP framework that applies paragraph-level classification via LLaMA 3.1 orchestrated through LangGraph. This modular approach enhances interpretability, enables type enforcement through structured parsing, and facilitates large-scale speech annotation with transparent logic. We apply this pipeline to over 1,200 Federal Reserve speeches and investigate whether thematic emphasis—quantified as the relative share of inflation, employment, or other topics—predicts abnormal changes in 10-year Treasury yields.

Literature Review. A growing body of research explores the relationship between central bank communication and financial markets. Early work relied on dictionary-based sentiment scoring of earnings calls and regulatory filings, such as the Loughran-McDonald dictionary tailored to financial texts (6). Subsequent studies applied topic modeling and textual regressions to FOMC minutes and policy speeches to capture the thematic evolution of central bank narratives (5). More recent developments have used transformer-based models to detect changes in tone and transparency in ECB and Federal Reserve communication (4), and to link central bank sentiment to short-term equity volatility and credit spreads.

However, these approaches often suffer from key limitations. Many rely on bag-of-words or TF-IDF representations, which ignore word order, contextual meaning, and structural cues. Document-level aggregation blurs intra-textual variation, making it difficult to isolate fine-grained signals. In particular, most existing NLP pipelines lack paragraph-level granularity and do not harness the compositional or state-aware reasoning capabilities of modern large language models (LLMs). Furthermore, few studies operationalize communication as a time-localized signal within an event-study design, especially in the context of long-term fixed income markets such as Treasury yields.

Research Gap and Contribution. This paper addresses two underexplored challenges in the literature. First, we examine how thematic *emphasis*—measured as the paragraph-level proportion of inflation, employment, or other topics—affects cumulative abnormal yield changes (CAY) around speech dates. While prior studies focus on short-term equity or currency responses to news or sentiment shocks, the bond market's reaction to narrative framing remains less well understood. This is especially relevant for long-term yields, which embed expectations about future policy paths and macroeconomic outlooks.

Second, we introduce a novel agentic NLP pipeline that applies paragraph-level tagging using LLaMA 3.1 orchestrated through LangGraph. This architecture enables modular processing, conditional logic, and structured validation—addressing common shortcomings in interpretability and reproducibility. Our framework allows for scalable, high-resolution annotation of economic texts and sets the stage for more robust downstream modeling. We apply this to over 1,200 speeches by Federal Reserve policymakers and evaluate whether thematic emphasis has predictive power over bond yield movements after controlling for the federal funds rate.

Motivation. Trillions of dollars in U.S. Treasury securities are influenced by perceived signals in Federal Reserve communications. Understanding how thematic framing—e.g., focus on inflation versus employment—affects market expectations has implications for bond pricing, risk premiums, and investor behavior. Moreover, advances in large language models (LLMs) now enable scalable analysis of qualitative narratives at paragraph-level resolution. This presents a unique opportunity to augment human interpretation with AI-driven insight and promote more transparent, data-backed economic forecasting.

Objectives. This paper pursues two goals: (1) to assess whether narrative emphasis in Powell's speeches influences 10-year Treasury yields via an event-study and regression framework, and (2) to evaluate the efficacy of agentic AI—via LangGraph and LLaMA 3.1—for structured economic text analysis. Our results contribute to emerging literature on central bank NLP and agent-based LLM orchestration, while also laying groundwork for richer multi-agent financial research.

2 Data Collection and Annotation

I scraped over 1,200 speeches from the Federal Reserveś website using a Selenium-based pipeline. Each speech was stored in PostgreSQL with metadata (ID, date, speaker, title, URL, content). A LangGraph pipeline orchestrates LLaMA 3.1 agents to classify each paragraph into one of three themes: inflation, employment, or other. There are two agents involved in this basic setup. First, one is to call LLM to annotate each paragraph of speech based on whether it emphasizes inflation, employment, or any other topic. The other agent summarizes this result of the annotation of each paragraph for each speech. The following are the components of the framework.

2.1 Data Scraper

We developed a Python-based scraper to extract speech texts from the Federal Reserve's official website[1] for the period 2006 through May 2025. The scraper is built using `Selenium`, a browser automation tool, to handle dynamic JavaScript content and pagination across archive pages.

Selenium Configuration

The scraper uses headless Chrome with the following configurations:

- **Dynamic content handling:** Each page is fully loaded before DOM parsing via `time.sleep(2)`.
- **Pagination:** A loop clicks the "Next" button iteratively until no further pages are available.
- **Rate limiting:** Delays are added between page loads to prevent throttling or IP blocking.

Error Handling and Validation

To ensure data integrity:

- **Retries:** Each failed download (e.g., timeout, bad status code) is retried up to 3 times with exponential backoff.
- **Validation:** Each speech is checked for:
  - non-empty title, date, and speaker metadata,
  - valid URL format,
  - text content exceeding 100 words.
- **Deduplication:** Speeches are uniquely identified by URL hash and date.
  Duplicates are dropped at ingest time.

Database Schema

All speeches are stored in a PostgreSQL database using the following schema:

```
CREATE TABLE fed_speeches (
    id SERIAL PRIMARY KEY,
    title TEXT NOT NULL,
    date DATE NOT NULL,
    speaker TEXT,
    url TEXT UNIQUE NOT NULL,
    text TEXT NOT NULL,
    created_at TIMESTAMP DEFAULT NOW()
);

CREATE TABLE fed_speech_analysis (
    id SERIAL PRIMARY KEY,
    speech_id INTEGER REFERENCES fed_speeches(id) ON D
    chair TEXT,
    emphasis JSONB, -- e.g.,
    --{"inflation": 0.3, "employment": 0.4, "other": 0
    tags JSONB,      -- list of {"paragraph": ..., "the
    created_at TIMESTAMP DEFAULT NOW()
);
```

This dual-table structure supports both raw text and structured NLP output. Indexes are created on `date` and `speaker` for efficient filtering.

Coverage

As of May 2025, the scraper has collected around 1,200 speeches, covering five Fed Chairs and spanning two decades of monetary communication. This dataset enables high-resolution analysis of narrative trends and their potential market effects.

[1] https://www.federalreserve.gov/newsevents/speeches.htm

2.2 *Annotation Agent*

This agent interacts with a language model (LLM) to classify the theme of speech paragraphs. It prepares the input prompt , calls the LLM, parses the response, and handles errors if parsing fails. The output is structured using Pydantic and stored as JSON emphasis scores.

2.3 *Emphasis Score Calculator Agent*

This agent uses the tags generated by the previous agent. We are using following approach to calculate the emphasis score:

Given n paragraphs in a speech, emphasis scores are computed as:

$$\text{Score}_{\text{theme}} = \frac{\text{Count}_{\text{theme}}}{n}$$

Similarly we calculate score for employment and other and store them in postgresdB corresponding to the original speech.

To address multicollinearity from the constraint that theme scores sum to one, we exclude the "other" category during regression analysis. This avoids the dummy variable trap and improves interpretability of model coefficients.

I also stored the tagged output in Postgresql database. The idea is to do prepare dataset on which analysis can be run independently. Since there are two components of the study: analysis and annotation I wanted to modify them independently for future improvement in the research. The speach annotation results are saved in format speech id, tags and emphasis.

The overall langraph architecture looks like following :

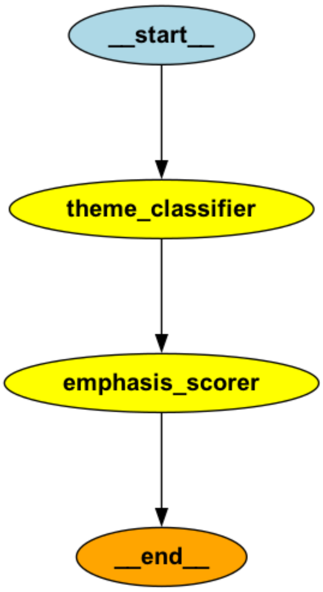

Figure 1. LangGraph structure For Annotation

figure 1 shows the design of LangGraph archirecture including two agents that are involved in the agentic flow.

2.4 *Event Study*

I downloaded fed data for 10 year treasury yield and forward fund rate data to contextualize the speeches data and analyse if the speech events with some specific focus has any impact on treasury yield.

3 Methodology

We tried to do event-based analysis of fed speeches. following are the frameworks and models used for the study.

3.1 *Agentic NLP Classification Pipeline*

Agentic AI refers to the design of autonomous, cooperating language model components (agents) that solve sub-problems within a structured workflow. Instead of monolithic prompt engineering, agents operate within a predefined topology—typically a graph—to accomplish tasks like extraction, classification, summarization, or validation. This structure improves modularity, reusability, and traceability (2).

We adopt the LangGraph framework for orchestrating a multi-agent workflow over unstructured central bank speech texts. The advantages of such orchestration over sequential or naive looping include:

- **State tracking:** Intermediate results such as paragraph tags and parser output are carried across nodes with type guarantees (via Pydantic).
- **Conditional logic:** Agents can conditionally reprocess or retry malformed generations, ensuring robustness in long-document inference.
- **Parallel execution:** Paragraphs are split and routed to classifier nodes in parallel, drastically improving throughput for large corpora.

3.1.1 *Agent Topology*

Our LangGraph configuration includes the following roles:

- **Theme Classifier Agent:** Splits speeches into atomic paragraph units, preserving semantic cohesion.Uses a LLaMA 3.1 model to assign each paragraph one of three themes: `inflation`, `employment`, or `other`. Enforces output format consistency using a Pydantic schema and handles incomplete or malformed LLM output.
- **Emphasis Calculator Agent** aggregates themes of all the paragraphs and calculates a combined emphasis score.

Each paragraph is passed independently to the classifier agent. LangGraph's state object accumulates the results, allowing us to derive per-speech emphasis vectors. This ensures both scalability and reproducibility.

3.1.2 *Comparison to Traditional Pipelines*

Traditional NLP pipelines for financial text analysis often rely on:

1. Dictionary-based tagging (e.g., LIWC, Loughran-McDonald),
2. Topic modeling (e.g., LDA) which assumes document-level distributions, or
3. Single-pass LLM summarization prompts which lack modular control.

In contrast, our agentic setup allows for:

- Fine-grained control over each processing step,

- Easy insertion of future agents (e.g., tone detector, coreference resolver),
- Debuggability: faulty generations are isolated and handled via retry logic,
- Multi-agent future extensions like role-specific disagreement analysis or speech rewriting debates.

3.1.3 *Extensibility for Economic Research*

The modular graph-based structure can support downstream tasks like:

- Automatic identification of narrative shifts across economic cycles,
- Extraction of forward guidance cues for predictive modeling,
- Cross-agent dialogue simulation (e.g., a "market analyst agent" responding to a "Fed speech agent").

Such a framework goes beyond classification towards interpretable, extensible systems for economic understanding. This reflects a broader trend toward agent-based reasoning in LLM applications to finance and policy.

3.2 *Event Study Framework*

We adopt a classic event-study design (7). The estimation window is [-120, -10] days before each speech; the event window is [0, +1] days. Abnormal yield changes (AY) and cumulative abnormal yields (CAY) are calculated as deviations from the estimation-period average yield change. The analysis was repeated for different estimation windows, event windows and results were analyzed for each. Mostly the conclusion of analysis across different windows parameters remains same. As an extension of the study we plan to analyze if there are going to be any window which due to some specific reason might have different results.

Let Y_t represent the 10-year Treasury yield on trading day t. Define the daily yield change as:

$$\Delta Y_t = Y_t - Y_{t-1}$$

Let $\mu_{\Delta Y}$ denote the average daily yield change estimated over the pre-event estimation window $[t-120, t-10]$. Then the **Abnormal Yield (AY)** on event day t is computed as:

$$AY_t = \Delta Y_t - \mu_{\Delta Y}$$

The **Cumulative Abnormal Yield (CAY)** over an event window of $[t, t+1]$ is defined as the sum of the abnormal yields:

$$CAY = \sum_{i=t}^{t+1} AY_i = (Y_t - Y_{t-1} - \mu_{\Delta Y}) + (Y_{t+1} - Y_t - \mu_{\Delta Y})$$

3.3 *Regression Model*

We regress CAY on emphasis scores and the effective federal funds rate (FFR).

The Federal Funds Rate (FFR) serves as the primary instrument of U.S. monetary policy and directly influences the short end of the yield curve. Movements in long-term Treasury yields, such as the 10-year note, reflect both the current policy stance and market expectations of future policy changes (9). Including the FFR as a control variable ensures that the regression model accounts for the baseline monetary conditions under which speeches are delivered.

Moreover, central bank speeches often reaffirm or deviate from prior policy signals, so omitting the actual FFR could lead to omitted variable bias. This is especially important given that our dependent variable, Cumulative Abnormal Yield (CAY), reflects short-term deviations from expected yield paths. By controlling for the FFR, we isolate the marginal impact of narrative emphasis from the mechanical effect of policy rate changes.

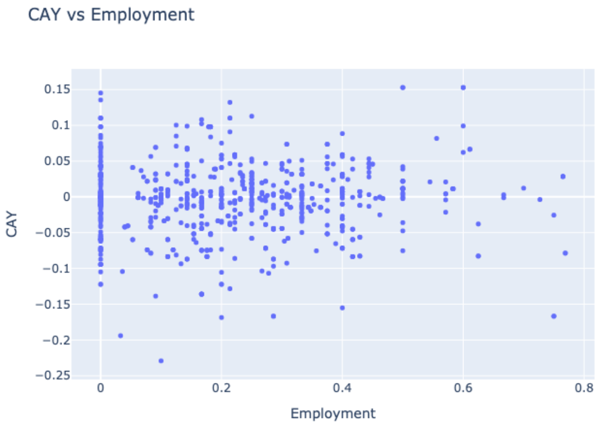

Figure 2. CAY vs employment graph.

Figure 2 shows the relation between CAY and employment. Clearly there is no linear relation between these two variables.

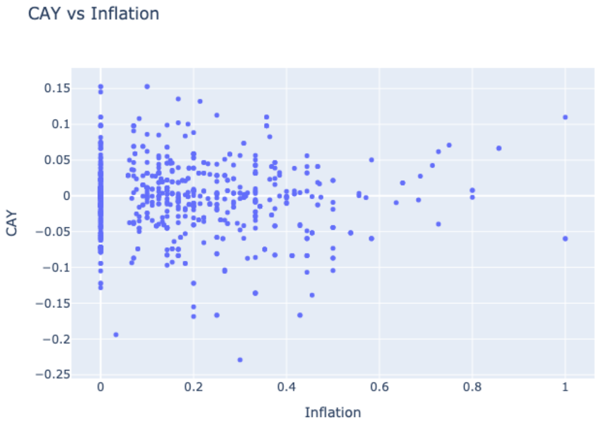

Figure 3. CAY vs inflation graph.

Figure 3 shows the relation between CAY and inflation.Clearly there is no linear relation between these two variables either.

Employment and inflation variables don't show a linear relation with CAY though. The distribution looks scattered and CAY values are centered around zero. However I am running regression model to study the combined effect of variables and keep open the possibility of modifying this model further by introducing new variable.

$$\mathrm{CAY}_t = \beta_0 + \beta_1 I_t + \beta_2 E_t + \beta_3 O_t + \beta_4 \mathrm{FFR}_t + \epsilon_t \qquad (1)$$

where I_t, E_t, and O_t denote emphasis on inflation, employment, and other themes respectively.

Variance Inflation Factor (VIF) diagnostics reveal severe multicollinearity among the emphasis score variables. Inflation, employment, and other emphasis scores yield VIF values exceeding 10^5, far above the common threshold of concern (typically 5–10) (8). This is expected because the three proportions sum to 1 by construction, creating perfect linear dependence. As a result, the regression suffers from unstable coefficient estimates and inflated standard errors.

The Federal Funds Rate, by contrast, shows a low VIF of 1.02, indicating it is not collinear with the emphasis scores. To mitigate this issue, one variable (typically "other") can be dropped to avoid the dummy variable trap, or an isometric log-ratio (ILR) transformation could be applied to convert the compositional data into orthogonal coordinates (3). To improve the regression analysis we removed the other tag from analysis and it improved the result of VIF test.

4 Results

4.1 *Descriptive Trends*

Exploratory data analysis shows that speeches varied thematically over time. A stacked area chart of quarterly emphasis reveals spikes in inflation-related content around 2022, coinciding with rising inflation expectations. This might have been in post pandemic market adjustment which give rise to worldwide inflation.

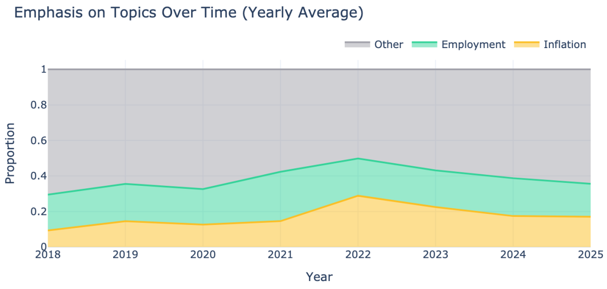

Figure 4. Quarterly thematic emphasis in Powell's speeches.

Figure 4 shows the evolving share of inflation, employment, and other content across Powell's tenure.

Figure 5 shows how different speakers in Powell's tenure emphasized on employment and inflation.

Figure 6 shows how the yield and fed fund varies over years.

Fig 7 shows that in different fund rate and treasury yield scenario what were the dominant topic. Concentration of inflation in upper right quadrant means inflation was the most discussed topic when fund rate and yield both were high.

The following heatmap shows pairwise correlations between emphasis scores (inflation, employment, other), Treasury yields, and the

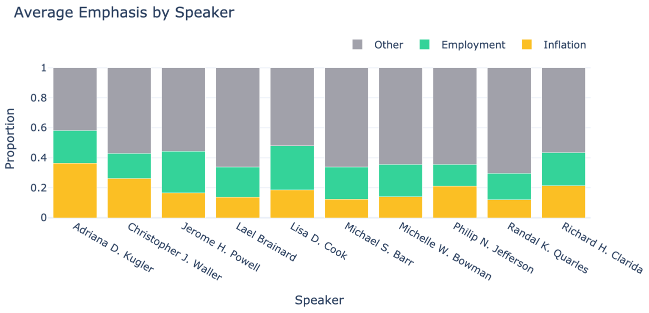

Figure 5. Distribution of thematic emphasis by speaker.

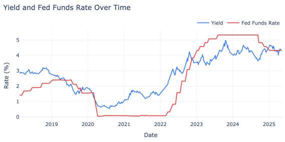

Figure 6. Treasury Yield rate and Fed Fund graph

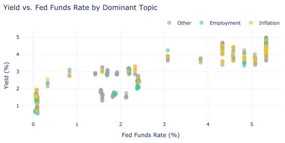

Figure 7. Treasury Yield and Fed Fund rate relation with dominant topics

federal funds rate. Since our regression model uses the 10-year Treasury yield (via CAY) as the dependent variable, it is important to assess multicollinearity among the independent variables.

Although inflation, employment, and other emphasis scores are inherently negatively correlated due to their additive nature (they sum to 1), we observe a relatively low correlation between these scores and external policy variables such as the federal funds rate. This supports their simultaneous inclusion in regression models. Low multicollinearity among independent variables enhances coefficient stability and interpretability (8).

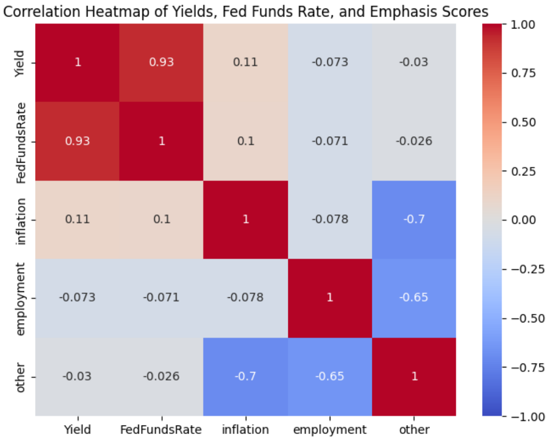

Figure 8. Correlation between emphasis scores and yield metrics.

4.2 *Regression Findings*

We estimate two ordinary least squared (OLS) models to assess whether thematic emphasis in Powell's speeches influences cumulative abnormal yields (CAY) on 10-year Treasuries. Both models control for the federal funds rate, which reflects prevailing monetary policy conditions.

4.2.1 *Full Model with Three Emphasis Themes*

Table 1 shows the regression results using all three emphasis variables: inflation, employment, and other. The regression includes 704 observations. While the model is statistically significant overall (F-statistic = 3.28, p = 0.011), the only significant predictor is the federal funds rate (p = 0.002). The model has an R^2 of 0.018, indicating limited explanatory power.

4.2.2 *Reduced Model with Two Emphasis Themes*

To address multicollinearity inherent in compositional data (emphasis scores summing to 1), we drop the "other" category and re-estimate the model. Table 2 reports results for 720 speeches.Variance inflation factors (VIFs) for the predictors fall near 1, confirming that collinearity is no longer a concern. The federal funds rate remains statistically significant (p = 0.003), while both emphasis variables are not.

Table 1. Regression Results using all emphasis scores and FFR

Variable	Coefficient	P-Value
Intercept	4.685	0.307
Inflation Emphasis	–4.692	0.306
Employment Emphasis	–4.676	0.308
Other Emphasis	–4.679	0.308
Federal Funds Rate	–0.003	0.002**

Table 2. Regression Results using inflation, employment, and FFR only

Variable	Coefficient	P-Value
Intercept	0.0059	0.214
Inflation Emphasis	–0.0145	0.171
Employment Emphasis	0.0037	0.740
Federal Funds Rate	–0.0028	0.003**

4.2.3 *Visual Diagnostics and Interpretation*

Figure 9 shows the distribution of CAY across high and low inflation emphasis speeches. Despite a slightly thicker right tail for high-emphasis speeches, the overall distributional overlap suggests no strong signal. Residual diagnostics (Q-Q plot, histogram, residual-vs-fitted) show non-normality and mild autocorrelation, limiting inferential precision.

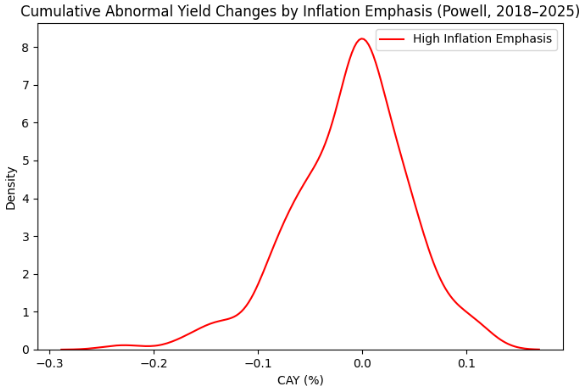

Figure 9. Cumulative abnormal yield changes for high vs. low inflation emphasis. Distributions overlap significantly.

4.2.4 *Summary*

These results suggest that Powell's thematic emphasis alone does not significantly influence long-term yield movements. However, the federal funds rate consistently emerges as a significant control, underscoring its relevance in event-driven interest rate studies. The findings reinforce the importance of including macroeconomic context when studying textual effects in finance.

5 Discussion

While our agent-based NLP pipeline captures narrative structure at scale, its predictive power for yield movement is limited. This may stem from:

- Emphasis scores being highly collinear.
- Market reactions being driven more by FOMC decisions than speech content.
- Equal weighting of paragraphs, regardless of length or position.

Future work may include sentiment analysis, hierarchical paragraph weighting, and expanding themes beyond the current taxonomy.

6 Conclusion

This paper demonstrates an end-to-end agentic NLP framework for analyzing central bank communications. By combining LangGraph orchestration, LLaMA 3.1 classification, and event-study regression, we quantify the influence of speech content on Treasury yields. While emphasis alone lacks predictive power in this model, the infrastructure offers a valuable foundation for future research in AI-driven macro-financial analysis.

Appendix: Model Diagnostics

Variance Inflation Factors (VIF)

To assess multicollinearity, we compute the Variance Inflation Factor (VIF) for each predictor in the regression model. Table 3 shows that all substantive predictors—`inflation`, `employment`, and `FedFundsRate`—have VIFs near 1.0, indicating minimal collinearity and stable coefficient estimation.

Table 3. VIF Values for Final Regression Model

Variable	VIF
Intercept	5.88
Inflation Emphasis	1.02
Employment Emphasis	1.01
Federal Funds Rate	1.01

The elevated VIF for the intercept (5.88) is not a concern. This value arises from the fact that the intercept absorbs shared variance due to centering and scaling of the explanatory variables, especially when compositional variables (like emphasis scores) are constrained to sum to a constant. Since the intercept is not a substantive variable, high VIF values for it do not threaten the stability or interpretability of the regression model (8).

Model Residual Diagnostics

We also assess two common OLS assumptions related to residual behavior:

- **Durbin-Watson Statistic:** 1.051 — indicates mild positive autocorrelation.
- **Jarque-Bera Test:** $p < 0.001$ — residuals deviate from normality.

While these diagnostics limit the precision of small-sample inference, they do not materially affect coefficient directionality or broader conclusions in a large-sample OLS framework. Future work may apply robust or bootstrapped standard errors to address these issues.

References

[1] Alan S Blinder, Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David-Jan Jansen, ‘Central bank communication and monetary policy: A survey of theory and evidence’, *Journal of Economic Literature*, **46**(4), 910–945, (2008).

[2] LangGraph Developers. Langgraph: Multi-agent orchestration framework for llms. https://github.com/langchain-ai/langgraph, 2024. Accessed: 2024-05-15.

[3] Juan J. Egozcue, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras, and Carles Barceló-Vidal, ‘Isometric logratio transformations for compositional data analysis’, *Mathematical Geology*, **35**(3), 279–300, (2003).

[4] Michael Ehrmann and Jonathan Talmi, ‘Semantic similarity in central bank communication and market volatility’, *Journal of Monetary Economics*, **114**, 262–277, (2020).

[5] Matthew Gentzkow, Bryan Kelly, and Matt Taddy, ‘The use of text as data’, *Journal of Economic Literature*, **57**(3), 535–574, (2019).

[6] Tim Loughran and Bill McDonald, ‘When is a liability not a liability? textual analysis, dictionaries, and 10-ks’, *The Journal of Finance*, **66**(1), 35–65, (2011).

[7] A. Craig MacKinlay, ‘Event studies in economics and finance’, *Journal of Economic Literature*, **35**(1), 13–39, (1997).

[8] Douglas C. Montgomery, Elizabeth A. Peck, and Geoffrey G. Vining, *Applied Regression Analysis and Generalized Linear Models*, SAGE Publications, Thousand Oaks, CA, 5th edn., 2021.

[9] Michael Woodford, *Interest and Prices: Foundations of a Theory of Monetary Policy*, Princeton University Press, Princeton, NJ, 2005. See Chapter 4: Monetary Policy in the Short Run.